REVISED DEPTH MAP ESTIMATION FOR MULTI-VIEW STEREO

Yao Yao, Hao Zhu, Yongming Nie, Xiaoli Ji, Xun Cao

Nanjing University
School of Electronic Science and Engineering
No.163, Xianlin Avenue, Nanjing, China

ABSTRACT

Optical flow estimation is one of the popular methods to obtain the depth maps in multi-view stereo due to its high accuracy and robustness. In traditional optical flow estimation, the energy function contains three assumptions: intensity constancy assumption, gradient constancy assumption, and global smoothness assumption. In this work, we propose a local smoothness assumption to constrain the optical flow disparity in neighboring pixels. We first study the new smoothness term and its corresponding energy function, and present a practical iteration approach to minimize the energy function. Later we apply this new estimation method to the multi-view stereo system and obtain the depth maps of different image pairs. Our results demonstrate the good performance of the algorithm in acquiring smoothing surface when comparing to the traditional methods.

Index Terms— Optical flow estimation, local smoothness assumption, multi-view stereo.

1. INTRODUCTION

Among all the methods to reconstruct the real-world object, multi-view stereo draw people's attention because of its excellent performance and potential applications in many fields, including industrial modeling, outdoor scene reconstruction [17, 27], film industrial [2, 4], 3D television [9, 8, 7] and the conceptual teleconference. Many efficient algorithms have been developed to solve this problem. According to the taxonomy of Seitz [25], these methods could be categorized into the following classes: 3D volumetric approaches [20, 26], surface evolution techniques [10, 24], feature extraction and expansion algorithms [13, 14, 11], and depth map based methods [3, 12, 23, 22].

Depth map based methods are the most popular methods due to the top-ranking results they produced on standard evaluation tests [1]. Generally these methods have two separate processing stages: depth map estimation, in which each view point's depth map is generated from a binocular stereo, and 3d points merging, that merges the depth maps into one 3d point cloud and produces the 3D model. The two-stage pipeline gives researchers great flexibility in choosing processing approaches, and the adaptability makes easy transplantation between different algorithms. In the depth map based methods, the key factor towards a high quality result is the estimation of the depth maps of two nearby images. Efforts to attain the high quality depth map of each view point have been made by researchers for years, including window based matching of [3, 12, 23, 6], surface-consistency metric of [28], and photometric stereo methods [16, 21]. However, these methods, as well as feature matching [11, 15], endeavor to rectify distorted matching windows or search the possible slanted angles for accuracy matching because of the distorted matching windows caused by slanted projection of surface patches. Optical flow estimation through the variational method is another common-used solution. We choose to acquire the depth map using this method in our work. Modifications to the energy function was made while applying the local smoothness assumption we propose.

Traditionally the target energy function of the optical flow model consists of three parts: an intensity constancy term concerning brightness consistency, a gradient consistency term for adjustment to brightness change and a global smoothness term to smooth the surface. However, it is not easy to find the balance between preserving the model details and ensuring the smoothness because of the knotty trade-off. In this paper, we propose the local smoothness assumption. Different from the traditional smoothness assumption which consider the displacement of pixels as the same in the adjacent area, the local smoothness assumption treats small areas on the object's surface as a plane. This assumption introduces a second spatial gradient term in the energy function and positively affects the quality of final reconstructed model (see picture 5). Corresponding details of the assumption are described in Section 2.

The organization of the paper is arranged as follow: In the next section, the new smoothness assumption and its corresponding local smoothness term are introduced. A numerical method which could practically solve the target energy minimization problem is also described in the next section. Section 3 discuss the reconstruction pipeline and mainly focus on the normalized-cross correlation method used in 3d point cloud merging. Compared with results using tradition-

978-1-4799-8023-9/14/$31.00

ally optical flow methods, the experimental results are given in Section 4.

2. DEPTH ESTIMATION WITH LOCAL SMOOTHNESS ASSUMPTION

2.1. Traditionally Variational Model

Before introducing the smoothness assumption, we studied the variational model used for common optical flow estimation. The old model includes three assumptions to constrain the target energy function.

Intensity constancy assumption. The assumption assumes the matching pixels in different image views share the same brightness, and will not be changed by displacement [18]. The mathematical expression is described as follow:

$$I(x, y, t) = I(x + u, y + v, t + 1) \tag{1}$$

Where I denotes the brightness of image pixels, and vector $\omega = (u, v, 1)$ represents the displacement between image at time t and image at time $t + 1$.

Gradient constancy assumption. The gradient constancy assumption regulates the brightness consistency in different images, however, do not take the changes of light condition into consideration. The gradient constancy assumption was introduced to solve this problem. It focuses on the gradient value of the brightness, which could tolerant slightly brightness change.

$$\nabla I(x, y, t) = \nabla I(x + u, y + v, t + 1). \tag{2}$$

Here the spatial gradient is denoted as $\nabla = (\partial_x, \partial_y)$. When dealing with the translatory motion situation, the gradient constancy assumption is particularly helpful. In contrast, intensity constancy constraint could be better suited for more complicated motion patterns.

Smoothness assumption. The two assumptions mentioned above only consider about the constraints on individual pixel, but fail to take the interaction between pixels into consideration. In fact, the surface of real world object satisfy the spatial continuity. Such condition limits the displacement of neighboring pixels to be less different from each other, which yields to:

$$\nabla u(x, y, t) = 0, \nabla v(x, y, t) = 0 \tag{3}$$

2.2. Local Smoothness Assumption

In order to constrain the spatial continuity the spatial derivatives of the displacement are set to be zero. However, this constraint, focuses on global smoothness rather than local regions, could raise knotty trade-off between smoothness and details. In this work, we assume the small area on object's surface as a plane (See figure 1). This constrain, not a strong regulation toward neighboring pixel, still holds the function of smoothing object's surface.

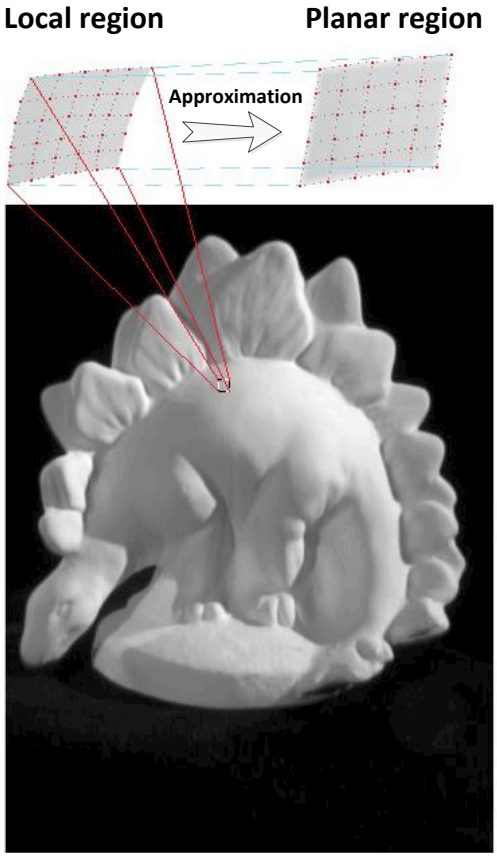

Fig. 1. Schematic of local smoothness assumption. A small region on the object's surface is considered as a planar region in order to achieve the local smoothness.

Suppose P is one point in plane S, then P could be expressed as $P = \lambda\hat{a} + \mu\hat{b}$, where $\hat{a}$, $\hat{b}$ are two base vectors of plane S, and λ, μ are two arbitrary constants. We define the projection points of point P in the neighboring cameras as p_l and p_r. p_l and p_r could be calculated for the calibrated cameras:

$$\begin{aligned} p_l &= P_{CL}P = P_{CL}(\lambda\hat{a} + \mu\hat{b}) \\ p_r &= P_{CR}P = P_{CR}(\lambda\hat{a} + \mu\hat{b}) \end{aligned} \tag{4}$$

Here, P_{CL} and P_{CR} refer to the Projection matrix of the left camera and the right camera. The disparity of the two pixels could be calculated: $D = (P_{CR} - P_{CL})(\lambda\hat{a} + \mu\hat{b})$. Disparity in this case is a linear function of its position (λ, μ), which leads to the following constraint on the pixel's displacement between two images:

$$\nabla^2 u = 0, \nabla^2 v = 0. \tag{5}$$

This constraint could result in second derivative term in the target energy function from the simple linear transforma-

tion. Compared to the first order derivative term caused by the traditional smoothness assumption, the second derivative term focus more on the local area of the pixel, which could be used as fine adjustment between smoothness and details.

2.3. Target Energy Function

According to the intensity constancy assumption, gradient constancy assumption, smoothness assumption and local smoothness assumption, the target energy function could be expressed from a linear combination of these four constraints. The adjusted energy function is:

$$E(\omega) = E_{Data}(\omega) + \alpha E_{Smooth}(\omega) \tag{6}$$

where

$$\begin{aligned} E_{Data}(\omega) = \int_{\omega} \phi_D(|I_r(p + d(\omega)) - I_t(p)|^2 \\ + \gamma|\nabla I_r(p + d(\omega)) - \nabla I_t(p)|^2)dxdy \end{aligned} \tag{7}$$

and

$$\begin{aligned} E_{Smooth}(\omega) &= \int_{\omega} \phi_S(|\nabla\omega|^2 + \beta|\nabla^2\omega|^2)dxdy \\ &= \int_{\omega} \phi_S(|\nabla u|^2 + |\nabla v|^2 \\ &\quad + \beta(|\nabla^2 u|^2 + |\nabla^2 v|^2))dxdy \end{aligned} \tag{8}$$

Here I_r and I_t refer to the reference image and target image, respectively. The data term contains two parts: the first part assumes the brightness consistency in each view point and the second part ensures the spatial gradient consistency. The introduction of gradient term ∇I makes the approach more robust to varying illumination and better preserves the edge of the object. Since some defeats such as noises, brightness changes and occlusions always make the data constraint inaccurate, two adjusted smoothness terms are introduced. Used to penalize the total variation of the flow field, the first smoothness term focuses on displacement consistency of nearby pixels. The second term focus on the smoothness of local areas.

To now the problem of finding the variables u, v that minimize the target variational energy function can be converted to the minimization of Euler-Lagrange equation. For simplicity in later explanation, we use the following abbreviations:

$$\begin{gathered} I_x := \partial_x I_r(p + d(\omega)), I_{xy} := \partial_{xy} I_r(p + d(\omega)), \\ I_y := \partial_y I_r(p + d(\omega)), I_{yy} := \partial_{yy} I_r(p + d(\omega)), \\ I_z := I_r(p + d(\omega)) - I_t(p), I_{xx} := \partial_{xx} I_r(p + d(\omega)), \\ I_{xz} := \partial_x I_r(p + d(\omega)) - \partial_x I_t(p), \\ I_{yz} := \partial_y I_r(p + d(\omega)) - \partial_y I_t(p) \end{gathered} \tag{9}$$

The corresponding Euler-Lagrange equation of the energy function could be expressed as:

$$\begin{aligned} &\phi_D^{'}(I_z^2 + \gamma(I_{xz}^2 + I_{yz}^2))(I_z I_x + \gamma I_{xz} I_{xx} + \gamma I_{yz} I_{xy}) \\ &- \alpha div(\phi_S^{'}((|\nabla u|^2 + |\nabla v|^2 + |\nabla^2 u|^2 + |\nabla^2 v|^2)\nabla u)) = 0 \end{aligned} \tag{10}$$

$$\begin{aligned} &\phi_D^{'}(I_z^2 + \gamma(I_{xz}^2 + I_{yz}^2))(I_z I_x + \gamma I_{xz} I_{xx} + \gamma I_{yz} I_{xy}) \\ &- \alpha div(\phi_S^{'}((|\nabla u|^2 + |\nabla v|^2 + |\nabla^2 u|^2 + |\nabla^2 v|^2)\nabla v)) = 0 \end{aligned} \tag{11}$$

2.4. Numerical Method

Like other multi-view stereo works that pixel matching is conducted under the epipolar constraint, the minimization only concerns about the points on the epipolar line. Using this constraint, we could reduce the searching space from 2D to 1D, which greatly reduces the computational cost and total running time.

The Euler-Lagrange is a nonlinear equation for the variables $\omega(x, y) = (u(x, y), v(x, y))$, and the work of minimizing the energy function is one of the most important and time consuming step in the processing pipeline. In order to fit the minimization into a linear approach, we use a practical iteration method to reduce the nonlinear minimization problem to a linear minimization problem, which has been surveyed in detail in Thomas's work [5]. The fixed point iteration is divided into an outer iteration and an inner iteration, which deal with the nonlinearity in updated ω and function $\phi^{'}$, respectively.

A coarse-to-fined strategy is also applied in the minimization process. Here an arbitrary down-sampling factor of the pyramid structure η is set between zero and one. The multiple starting scales (MSS) framework is used as the initialization of the whole pyramid algorithm, and the outcome depth map of each pyramid level will be used as the initial condition of pixel matching in next level.

3. POINT CLOUD SYNTHESIS

Due to noises, brightness changes and other factors that would affect the depth map, the 3D point cloud combined from depth maps must be refined before meshing. Normalized cross correlation (NCC) filter and other constraints are introduced in our work to remove the outlier. The refined 3D point cloud is then meshed using Poisson reconstruction algorithm. Figure 2 illustrate the whole processing pipeline.

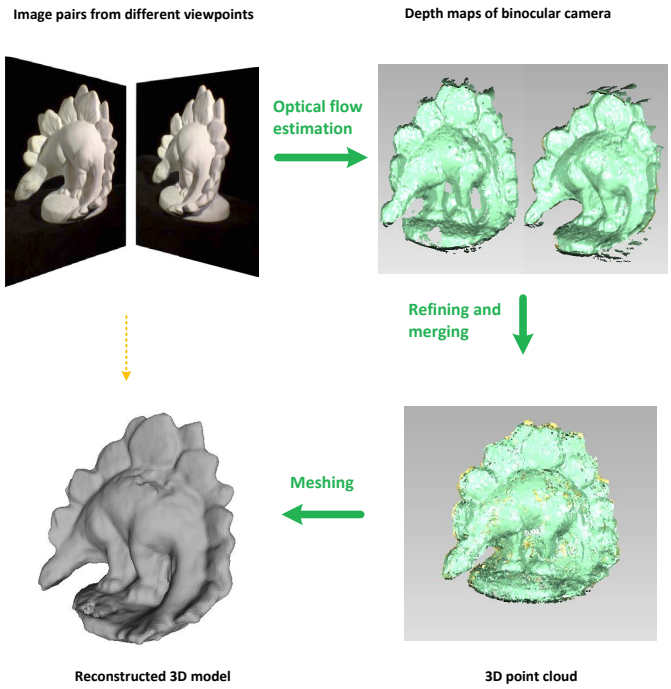

Fig. 2. Processing pipeline of our work. In the first step, depth maps are calculated through optical flow estimation. Later depth maps from different view point are synthesized to a 3D point cloud and then meshed to a watertight stereo.

3.1. Normalized Cross Correlation

To describe the fidelity of specific depth map candidates, we calculate the NCC value between the corresponding patch of a 3D point in different images. In this case, NCC indicates higher reliability while its value is closer to 1. NCC matching in image space could be defined as follow:

$$NCC(p_i, s) = \frac{\sum_{j=1}^{N^2}(n_j - \bar{n}) \cdot (f_s(n_j) - f_s\bar{(n)})}{\sqrt{\sum_{j=1}^{N^2}(n_j - \bar{n})^2} \cdot \sqrt{\sum_{j=1}^{N^2}(f_s(n_j) - f_s\bar{(n)})^2}} \quad (12)$$

n_j refers to local region of size $N \times N$ in the target image, p_i is the pixel in the primary view and s is the index of depth map. Generally we only select the biggest NCC value in all depth map candidates as the NCC value of pixel p_i.

3.2. Refinement and Poisson Reconstruction

To remove the outliers in 3D point cloud so as to improve the accuracy of the final point cloud, we regard points with neighbor number less than $0.5n$ as outliers. Here n is the average neighbor number. The normals of the points are also used as a constraint: only points whose angle between normals and view-vector are larger than 45 degree are retained. Visual hull constraint is then employed again to remove those obvious points outside the model. Figure 3 shows the result of 3D point merging and refinement.

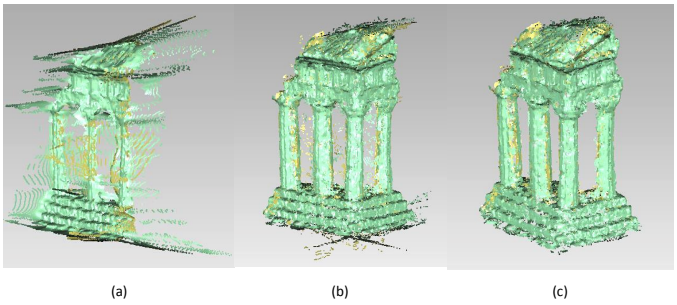

Fig. 3. The raw depth map is shown in picture (a). Picture (b) are the point cloud filtered through NCC filter. Picture (c) show the final point cloud used for meshing.

Finally, the point cloud is meshed through Poisson surface reconstruction technique [19] to obtain a watertight model. The final model bears the robustness to both noise and non-uniform sampling rate.

4. EXPERIMENTAL RESULTS

4.1. Implementation

Figure 4 shows the reconstruction results of *dinoSparseRing* and *templeSparseRing* from Middlebury dataset [1]. In the experiment we test the *dinoSparseRing* and *templeSparseRing* datasets. The main parameters used in the experiment are listed in table 1. For comparison, we set the weight value β as zero to represent the situation of traditionally optical flow estimation method. The pyramid level used in both models is 15.

Table 1. Main parameters used in the experiments

Parameters	Value
Smooth Parameter α	10
Weight Factor γ	90
Local Smoothness Parameter β	1
Outer Fixed Point Iteration	5
Outer Fixed Point Iteration	5
Inner Fixed Point Iteration	2
SOR Iteration	10
Down-sampling Factor η	0.9

The main computational time is spent on the step of optical flow estimation which is greatly related to the resolution

of the images. Middlebury datasets have a low resolution of 640×480, and this make the computation time satisfactory for model reconstruction. The whole processing time of our algorithm is about 25 minutes.

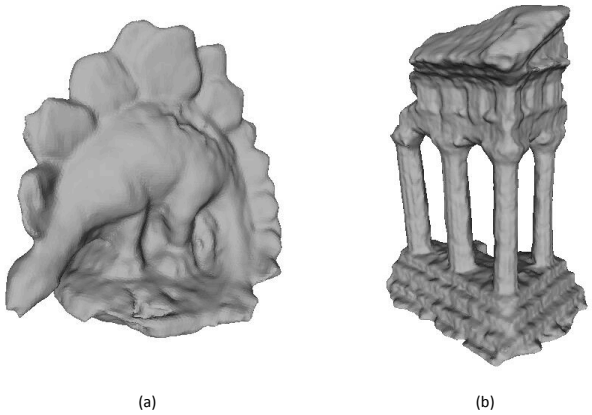

Fig. 4. Reconstruction results of *dinoSparseRing* and *templeSparseRing* from Middlebury College's dataset

4.2. Reconstruction Results and Comparison

Figure 5 (a) shows the reconstruction model using traditionally optical flow estimation algorithm. Figure 5 (b) is the result of our algorithm with $\beta = 1$. Green circle regions shows the refined surface while the region of red circles represent the coarse surface reconstructed from traditional method. Figures 5 (c) and (d) are the reverse sides of (a) and (b) respectively.

The results of our algorithm show the good performance in surface smoothing. With the augmentation of β, the surface of the object become smoother. When pyramid level, smoothness parameter α and β are set as 15, 10, and 1.0, the algorithm produces a good reconstruction model. In our experiment, the value of β between 1 to 3 produces better reconstructed model than other value. However, as the case in previous energy function that trade off could be aroused by smoothness term, setting a larger β to the local smoothness term would cause over smoothness. Local smoothness term could be used as a fine adjustment variable combined with the traditional smoothness term.

5. CONCLUSION

We propose the local smoothness assumption while estimating the optical flow between two nearby cameras, and apply the new algorithm to multi-view stereo. The assumption shows its good performance in smoothing object's surface when combining with the traditional smoothness assumption. However, due to over smoothing, the local smoothness parameter should be chosen appropriately. There are rooms for

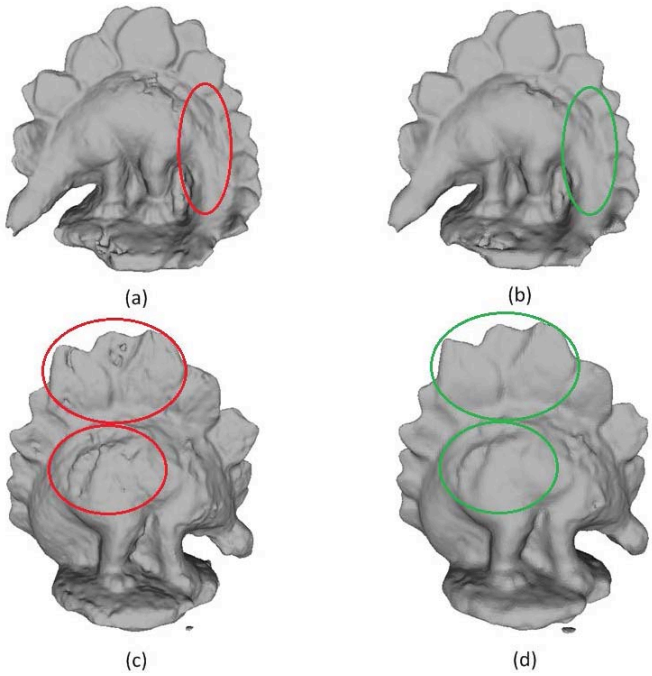

Fig. 5. Influence of local smoothness term on experimental result. Picture (a1) and (b1) denote the traditional reconstruction model, while (a2) and (b2) are the results from new algorithm with local smoothness term.

improving the performance of our algorithm in future works. Like many top-ranking reconstruction results in Middlebury dataset which use optical flow estimation methods, we could refine our reconstruction result through parameter adjustment. Meanwhile, the simple merging strategy in our processing procedure could be improved for a high quality 3D point cloud, and high-quality surface refinement technique could be applied to the reconstruction pipeline.

6. ACKNOWLEDGEMENT

The authors would like to thank the support of the National NSF of China grant No.61371166, No.61422107 and the NSF of Jiangsu Province, China grant No.BK20130583. This work is also supported by the Fundamental Research Funds for the Central Universities No.20620140395, No.20620140417.

7. REFERENCES

[1] http://vision.middlebury.edu/mview/.

[2] O. Alexander, M. Rogers, W. Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and P. Debevec. The digital emily project: Achieving a photorealistic digital actor. *Computer Graphics and Applications, IEEE*, 30(4):20–31, July 2010.

[3] D. Bradley, T. Boubekeur, and W. Heidrich. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[4] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. *ACM Transactions on Graphics (TOG)*, 29(4):41, 2010.

[5] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision-ECCV 2004*, pages 25–36. Springer, 2004.

[6] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision–ECCV 2008*, pages 766–779. Springer, 2008.

[7] Xun Cao, A.C. Bovik, Yao Wang, and Qionghai Dai. Converting 2d video to 3d: An efficient path to a 3d experience. *MultiMedia, IEEE*, 18(4):12–17, April 2011.

[8] Xun Cao, Yebin Liu, and Qionghai Dai. A flexible client-driven 3dtv systemfor real-time aquisition, transmission, and display of dynamic scenes. *EURASIP J. Appl. Signal Process.*, 2009:5:1–5:15, January 2008.

[9] Xun Cao, Yebin Liu, Xiangyang Ji, and Qionghai Dai. Vision field capture for advanced 3dtv applications. In *Visual Communications and Image Processing (VCIP), 2011 IEEE*, pages 1–4, Nov 2011.

[10] Carlos Hernndez Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367 – 392, 2004. Special issue on model-based and image-based 3D scene representation for interactive visualization.

[11] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1362–1376, Aug 2010.

[12] M. Goesele, B. Curless, and S.M. Seitz. Multi-view stereo revisited. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2402–2409, 2006.

[13] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S.M. Seitz. Multi-view stereo for community photo collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.

[14] M. Habbecke and L. Kobbelt. A surface-growing approach to multi-view stereo reconstruction. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[15] M. Habbecke and L. Kobbelt. A surface-growing approach to multi-view stereo reconstruction. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[16] C. Hernandez, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(3):548–554, March 2008.

[17] Vu Hoang Hiep, Renaud Keriven, Patrick Labatut, and J-P Pons. Towards high-resolution large-scale multi-view stereo. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1430–1437. IEEE, 2009.

[18] Berthold K. Horn and Brian G. Schunck. Determining optical flow, 1981.

[19] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006.

[20] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *Computer VisionECCV 2002*, pages 82–96. Springer, 2002.

[21] Jongwoo Lim, J. Ho, Ming-Hsuan Yang, and D. Kriegman. Passive photometric stereo from motion. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1635–1642 Vol. 2, Oct 2005.

[22] Yebin Liu, Xun Cao, Qionghai Dai, and Wenli Xu. Continuous depth estimation for multi-view stereo. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2121–2128, June 2009.

[23] P. Merrell, A Akbarzadeh, Liang Wang, P. Mordohai, J.-M. Frahm, Ruigang Yang, D. Nister, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.

[24] J.-P. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 822–827 vol. 2, June 2005.

[25] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528. IEEE, 2006.

[26] Son Tran and Larry Davis. 3d surface reconstruction using graph cuts with surface constraints. In *Computer Vision–ECCV 2006*, pages 219–231. Springer, 2006.

[27] H.-H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):889–901, May 2012.

[28] Chenglei Wu, Xun Cao, and Qionghai Dai. Accurate 3d reconstruction via surface-consistency. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, pages 1–4, May 2009.